

Rec'd PCT/PTO 21 SEP 2005

(12)特許協力条約に基づいて公開された国際出願

(19) 世界知的所有権機関
国際事務局



(43) 国際公開日
2004 年 7 月 15 日 (15.07.2004)

PCT

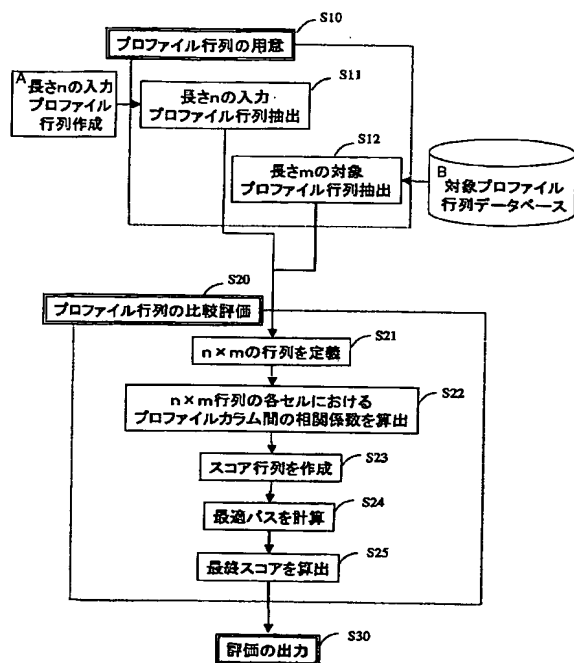
(10) 国際公開番号
WO 2004/059557 A1

- (51) 国際特許分類: G06F 19/00 (71) 出願人 (米国を除く全ての指定国について): 独立行政法人産業技術総合研究所 (NATIONAL INSTITUTE OF ADVANCED INDUSTRIAL SCIENCE AND TECHNOLOGY) [JP/JP]; 〒100-8921 東京都千代田区霞が関 1 丁目 3 番 1 号 Tokyo (JP).
- (21) 国際出願番号: PCT/JP2003/016982
- (22) 国際出願日: 2003 年 12 月 26 日 (26.12.2003)
- (25) 国際出願の言語: 日本語 (72) 発明者; および (75) 発明者/出願人 (米国についてのみ): 富井 健太郎 (TOMII, Kentaro) [JP/JP]; 〒135-0064 東京都江東区青海 2-4 1-6 独立行政法人産業技術総合研究所内 Tokyo (JP).
- (26) 国際公開の言語: 日本語
- (30) 優先権データ:
特願 2002-377704
2002 年 12 月 26 日 (26.12.2002) JP
特願 2003-406776 2003 年 12 月 5 日 (05.12.2003) JP
- (81) 指定国 (国内): AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM,

[続葉有]

(54) Title: SYSTEM FOR PREDICTING THREE-DIMENSIONAL STRUCTURE OF PROTEIN

(54) 発明の名称: タンパク質立体構造予測システム



S10...PREPARE PROFILE MATRIX
A...PREPARE INPUT PROFILE MATRIX OF LENGTH OF n
S11...EXTRACT INPUT PROFILE MATRIX OF LENGTH n
S12...EXTRACT OBJECT PROFILE MATRIX OF LENGTH m
B...OBJECT PROFILE MATRIX DATABASE
S20...COMPARE AND EVALUATE PROFILE MATRIX
S21...DEFINE n x m MATRIX
S22...CALCULATE COEFFICIENT OF CORRELATION BETWEEN PROFILE COLUMNS IN EACH CELL OF n x m MATRIX
S23...MAKE SCORE MATRIX
S24...CALCULATE OPTIMUM PATH
S25...CALCULATE FINAL SCORE
S30...OUTPUT EVALUATION

(57) Abstract: A system for evaluating the similarity between protein profile matrices, preferably usable for prediction of the three-dimensional structure of a protein. A profile matrix is composed of profile columns provided with the appearance probabilities of amino acids at the positions of the amino acid residues in a multiple alignment in which the amino acid sequences of relevant proteins are multiply arranged. The similarity evaluating system comprises (a) means for preparing two matrices, an input profile matrix and an object profile matrix, (b) means for calculating the coefficient of correlation between a profile column of the input profile matrix and a profile column of the object profile matrix for all or a part of the combinations of the profile columns, and (c) means for making a score matrix composed of the coefficients of correlation.

(57) 要約: タンパク質の立体構造予測に好適に使用できる、タンパク質プロファイル行列間の類似性評価システムを提供する。本発明は、タンパク質プロファイル行列間の類似性を評価するシステムであって、プロファイル行列は、関連する複数のタンパク質のアミノ酸配列を多重並置させたマルチプルアライメントにおいて、各アミノ酸残基位置におけるアミノ酸種毎の出現確率を備えたプロファイルカラムの群から構成され、(a) 入力プロファイル行列と、対象プロファイル行列の2つのプロファイル行列を用意する手段と、(b) 前記入力プロファイル行列の各プロファイルカラムと、前記対象プロファイル行列の各プロファイルカラムとの間の相関係数を、各プロファイルカラムの全部又は一部の組合せについて算出する手段と、(c) 前記相関係数からなるスコア行列を作成する手段とを含む。

WO 2004/059557 A1



HR, HU, ID, IL, IN, IS, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RU, SC, SD, SE, SG, SK, SL, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, YU, ZA, ZM, ZW.

FR, GB, GR, HU, IE, IT, LU, MC, NL, PT, RO, SE, SI, SK, TR), OAPI 特許 (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

添付公開書類:

— 国際調査報告書

(84) 指定国 (広域): ARIPO 特許 (BW, GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZM, ZW), ユーラシア特許 (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), ヨーロッパ特許 (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI,

2 文字コード及び他の略語については、定期発行される各 PCT ガゼットの巻頭に掲載されている「コードと略語のガイダンスノート」を参照。

明細書

タンパク質立体構造予測システム

5 技術分野

本発明は、タンパク質プロファイル行列間の類似性を評価するシステムに関するものであり、より詳しくは、タンパク質の立体構造予測に好適に使用されるタンパク質プロファイル行列間の類似性の評価システムに関する。

10 背景技術

自然界にあるタンパク質は進化の過程で選択され、特定の機能を発現するに至ったが、このタンパク質の機能はその立体構造に依存することが知られている。したがって、タンパク質の立体構造が予測できれば、その機能を予測することが可能となる。

- 15 従来、未だ何の知見も得られていないタンパク質を調べるに際し、既に立体構造が知られているタンパク質との類似性をコンピュータによって測定することにより、タンパク質の立体構造を推論ないし予測する手法が望まれていた。このような手法の1つとして、タンパク質プロファイル行列同士を比較する方法が、有力な手法として知られている (Rychlewski L, Jaroszewski L, Li W, Godzik A. Protein Sci (2000) Feb;9(2):232-41: 非特許文献1)。
- 20

- ここで、タンパク質プロファイル行列とは、関連するタンパク質 (タンパク質ファミリーなど) におけるアミノ酸種の出現頻度を、そのアミノ酸残基位置毎に数値化して行列としたものである。この行列は、通常、以下の手順で作成される。すなわち、まず、関連する複数のタンパク質のアミノ酸配列を多重並置させた
- 25 マルチプルアラインメントが与えられると、マルチプルアラインメントの各アミノ酸残基位置における20種のアミノ酸の各種類の出現数が計算される。続いて、これらの数を規格化することによって、出現確率に転換される。この時、与えられたマルチプルアラインメントに含まれるメンバー内での相互のアミノ酸配列類似性に応じた重みが考慮された上で出現数が補正され、プロファイル行列が作

成される。

ここで、マルチプルアラインメントとは、生物学的に相互に関連しあう複数のタンパク質のアミノ酸配列を、対応すると考えられるアミノ酸残基を揃えて並置したものをいう。マルチプルアラインメントは、例えば、ある一配列を入力値として、既存のプログラムであるPSI-BLAST(Altschul et al., Nucleic Acids Res. (1997) 25(17):3389-3402: 非特許文献2)を用いて、配列データベースに検索をかけることや、生物学的に相互に関連しあう複数のタンパク質のアミノ酸配列の一群を入力値として、これも既存のプログラムであるCLUSTALW(Higgins D., Thompson J., Gibson T. Thompson J.D., Higgins D.G., Gibson T.J. (1994). Nucleic Acids Res. 22:4673-4680: 非特許文献3)を用いることで容易に作成することができる。また、立体構造比較などの結果から作成することも可能である。

表1は、アミノ酸配列の長さ（アミノ酸残基数）がnであるタンパク質を基準として作成されたマルチプルアラインメントを模式的に示したものである。なお、表1中、第1列目は個々のタンパク質の名称であり、第1行目の「1～n」は、マルチプルアラインメントにおけるアミノ酸残基位置を示す。また、表1中のアルファベットはアミノ酸種を1文字標記したものである。

【表1】

	1	2	3	4	5	6	7	8	...	n
20807455/14-218	M	I	D	H	T	L	L	K	...	G
19551629/13-215	I	L	D	Y	T	L	L	G	...	A
16974933/15-229	L	M	D	L	T	T	L	N	...	A
16120769/20-234	L	M	D	L	T	T	L	N	...	A

表1の例では、例示されたアミノ酸残基位置のすべてにアミノ酸が配置されているが、アミノ酸残基位置に対応するアミノ酸残基がないとされた場合は、「・（ドット）」としてギャップを示すこともできる。表2は、表1で得られた長さがnであるマルチプルアラインメントにしたがって作成されたプロファイル行列を模式的に示したものである。表2中、第1列目はアミノ酸種（ギャップを含ん

でいてもよい) であり、第 1 行目の「1～n」は、プロファイル行列におけるアミノ酸残基位置を示す。

【表 2】

AA/Pos.	1	2	3	...	n
A	0.00	0.00	0.00	...	0.71
R	0.00	0.00	0.00	...	0.00
N	0.00	0.00	0.00	...	0.00
D	0.00	0.00	0.96	...	0.00
C	0.00	0.00	0.00	...	0.00
Q	0.00	0.00	0.00	...	0.00
E	0.00	0.00	0.04	...	0.00
G	0.00	0.00	0.00	...	0.29
H	0.00	0.00	0.00	...	0.00
I	0.29	0.29	0.00	...	0.00
L	0.41	0.29	0.00	...	0.00
K	0.00	0.00	0.00	...	0.00
M	0.29	0.41	0.00	...	0.00
F	0.00	0.00	0.00	...	0.00
P	0.00	0.00	0.00	...	0.00
S	0.00	0.00	0.00	...	0.00
T	0.00	0.00	0.00	...	0.00
W	0.00	0.00	0.00	...	0.00
Y	0.00	0.00	0.00	...	0.00
V	0.01	0.01	0.00	...	0.00

5

プロファイル行列中の各列は、関連する複数のタンパク質における、各アミノ酸残基位置の全アミノ酸種の確率分布を表すことになる。表 3 は、表 2 に示されたプロファイル行列のうち、残基位置が「2」であるプロファイルカラムを模式的に示したものである。

【表 3】

2
0.00
0.00
0.00
0.00
0.00
0.00
0.00
0.00
0.00
0.29
0.29
0.00
0.41
0.00
0.00
0.00
0.00
0.00
0.00
0.01

すなわち、表 2 で示されるプロファイル行列では、残基位置が 2 におけるアラニン (A) の補正された出現確率は 0. 0 0 であり、メチオニン (M) の補正された出現確率は 0. 4 1 ということになる。

従来、2つのプロファイル行列や2つのアミノ酸配列を比較及び／又は揃えるために、ダイナミックプログラミング (Needleman SB, Wunsch CD, J Mol Biol. (1970) Mar;48(3):443-53 : 非特許文献 4) が使用されてきた。アラインメントを作成する時に、比較される2つのアミノ酸配列や2つのプロファイル行列中のどの残基又はプロファイルカラムを対応付させるか (そこでは残基とギャップとの対応付も含まれる) 決定する必要があるが、その対応付のさせ方は非常に多数考えられる。ダイナミックプログラミングは、その中から類似性スコアが最大となるような対応付を自動的に効率良く見出すアルゴリズムである。そしてまた、その対応付の結果それ自体が最終的に得たいアラインメントである。

ダイナミックプログラミングでは、通常のアミノ酸配列比較の場合は、比較さ

れる2つのアミノ酸配列、および、比較したい2つのアミノ酸配列の各々の残基
 ペアに対する類似性スコア（類似の度合いを示す点数）から構成されるスコア行
 列、プロファイル行列比較の場合は、比較される2つの代表アミノ酸配列と、比
 較したい2つのプロファイル行列の、各々のプロファイルカラムのペアに対する
 5 類似性スコアから構成されるスコア行列の入力を要求する。これらを入力するこ
 とによって、ダイナミックプログラミングは、通常のアミノ酸配列比較の場合は
 、比較されるアミノ酸配列対のアラインメントとその最終スコア（類似性スコア
 が最大となるような最適パスを見つけることにより得られたスコア値）、プロフ
 ァイル行列比較の場合は、比較される代表アミノ酸配列のアラインメント、およ
 10 びその最終スコアを出力する。

したがって、ダイナミックプログラミングを使用する手法によりプロファイル
 行列を比較するためには、比較したい2つのプロファイル行列の類似性を精度よ
 く評価したスコア行列を作成する必要がある。

2つのプロファイル間の類似の程度を示すスコア行列を算出する方法として、
 15 Rychlewskiらが開発した手法が知られている（Rychlewski et al. (2000),
 9:p232-241）。これは、比較したいプロファイルカラムペア間の類似性スコアを
 、2つのプロファイルカラムを内積したものと定義づけて算出することにより、
 比較したい2つのプロファイル行列間のスコア行列を作成するものである。

たとえば、2つのプロファイル行列、 $X = x_1 x_2 \cdots x_p \cdots x_n$ （ただし、 x_p は
 20 アミノ酸残基位置 p におけるプロファイルカラム）および $Y = y_1 y_2 \cdots y_q \cdots y_m$
 （ただし、 y_q はアミノ酸残基位置 q におけるプロファイルカラム）が与えられ
 たとき、 n 行 m 列のスコア行列の要素である、類似性スコア D_{qp} （プロファイル
 カラム x_p およびプロファイルカラム y_q 間の類似性スコア）は、下記の式によっ
 て与えられる。

25 【数1】

$$D_{pq} = \sum_a^j x_{pa} y_{qa}$$

[式中、 $x_{p,a}$ = プロファイルカラム x_p の要素

$y_{q,a}$ = プロファイルカラム y_q の要素

j = プロファイルカラムの要素数 (通常 20) である。]

当該手法によれば、比較したい 2 つのプロファイルカラム間において、共にア
 ミノ酸置換が激しくない出現残基種が非常に限られている場合には、内積した値
 も高い数値となるため、高い類似性スコアが与えられる事になる。このように出
 現残基種が非常に限られておりアミノ酸変異が激しくない高度に保存されている
 残基位置は、生体内での機能的あるいは、物理化学的要請から高度に保存された
 箇所と考えられ、生物学的にも重要な位置であると考えられている。上記手法で
 は、このような領域はその類似性を精度良く評価することができると考えられる
 。

しかしながら、上記手法では、こうした出現残基種が限られた位置を精度良く
 評価することができる可能性があるものの、生物学的に重要な位置であっても、
 モチーフ内に存在する非保存位置や、タンパク質立体構造上露出していることが
 重要で極性のみが重大な意義を占める位置、あるいはその逆に埋没部分に位置し
 疎水性のみが保存されている位置など、アミノ酸置換が激しく生起していてもそ
 の置換パターンに共通性があると考えられるような領域に関して精度良く評価す
 ることができないという問題があった。

さらに、スコア行列の各要素 (類似性スコア) の平均値は負の値である事、標
 準偏差もほぼ一定値である事が望まれるため、類似性スコアに対して正規化処理
 を施さなければならず、煩雑であるという問題もあった。

従って、プロファイル行列間において、保存領域のみならず、非保存領域の類
 似性も評価できる、高精度かつ簡便な手法の開発が望まれていた。

【非特許文献 1】

Rychlewski L, Jaroszewski L, Li W, Godzik A. Protein Sci 2000 Feb;9(2):232-41

【非特許文献 2】

Altschul et al., Nucleic Acids Res. (1997) 25(17):3389-3402

【非特許文献 3】

Higgins D., Thompson J., Gibson T. Thompson J.D., Higgins D.G., Gibson

T. J. (1994). Nucleic Acids Res. 22:4673-4680

【非特許文献 4】

Needleman SB, Wunsch CD, J Mol Biol. 1970 Mar;48(3):443-53

5 発明の開示

本発明は、タンパク質の立体構造を予測するための、タンパク質プロファイル行列同士の類似性を評価するシステムを提供することを目的とする。

すなわち、本発明は、次のようなタンパク質プロファイル行列間の類似性評価システム、タンパク質立体構造の予測システム、コンピュータをそれらシステムとして機能させるためのプログラム、そのプログラムを記録したコンピュータ読み取り可能な記録媒体等を提供する。

(1) タンパク質の立体構造を予測するための、タンパク質プロファイル行列間の類似性を評価するシステムであって、

前記プロファイル行列は、関連する複数のタンパク質のアミノ酸配列を多重並置させたマルチプルアラインメントにおいて、各アミノ酸残基位置におけるアミノ酸種毎の出現確率を備えたプロファイルカラムの群から構成され、

前記類似性評価システムは、以下の手段：

(a) 立体構造を予測したいタンパク質を含む複数のタンパク質に基づいて作成される入力プロファイル行列と、立体構造が既知である複数のタンパク質に基づいて作成される対象プロファイル行列の2つのプロファイル行列を用意する手段と、

(b) 前記入力プロファイル行列の各プロファイルカラムと、前記対象プロファイル行列の各プロファイルカラムとの間の相関係数を、各プロファイルカラムの全部又は一部の組合せについて算出する手段と、

(c) 前記相関係数からなるスコア行列を作成する手段とを含むシステム。

(2) (1) 記載のシステムにより作成されたスコア行列を用いることを特徴とするタンパク質立体構造の予測システム。

(3) コンピュータを、タンパク質の立体構造を予測するためのタンパク質プロ

ファイル行列間の類似性を評価するシステムとして機能させるためのプログラムであって、

前記プロファイル行列は、関連する複数のタンパク質のアミノ酸配列を多重並置させたマルチプルアラインメントにおいて、各アミノ酸残基位置におけるアミノ酸種毎の出現確率を備えたプロファイルカラムの群から構成され、

前記類似性評価システムは、以下の手段：

(a) 立体構造を予測したいタンパク質を含む複数のタンパク質に基づいて作成される入力プロファイル行列と、立体構造が既知である複数のタンパク質に基づいて作成される対象プロファイル行列の2つのプロファイル行列を用意する手段と、

(b) 前記入力プロファイル行列の各プロファイルカラムと、前記対象プロファイル行列の各プロファイルカラムとの間の相関係数を、各プロファイルカラムの全部又は一部の組合せについて算出する手段と、

(c) 前記相関係数からなるスコア行列を作成する手段と

を含むプログラム。

(4) 上記(3)記載のプログラムを記録したコンピュータ読み取り可能な記録媒体。

(5) タンパク質プロファイル行列間の類似性を評価する方法であって、

前記プロファイル行列は、関連する複数のタンパク質のアミノ酸配列を多重並置させたマルチプルアラインメントにおいて、各アミノ酸残基位置におけるアミノ酸種毎の出現確率を備えたプロファイルカラムの群から構成され、

前記類似性評価方法は、以下のステップ：

(a) 入力プロファイル行列と、対象プロファイル行列の2つのプロファイル行列を用意するステップと、

(b) 前記入力プロファイル行列の各プロファイルカラムと、前記対象プロファイル行列の各プロファイルカラムとの間の相関係数を、各プロファイルカラムの全部又は一部の組合せについて算出するステップと、

(c) 前記相関係数からなるスコア行列を作成するステップとを含む方法。

(6) 前記対象プロファイル行列が、立体構造が既知である複数のタンパク質に基づいて作成されるプロファイル行列であり、前記入力プロファイル行列が、立体構造を予測したいタンパク質を含む複数のタンパク質に基づいて作成されるプロファイル行列である上記(5)記載の類似性評価方法。

- 5 (7) 上記(5)又は(6)で得られたスコア行列を用いることを特徴とするタンパク質立体構造の予測方法。

本発明により、タンパク質プロファイル行列間の類似性を簡便かつ精度よく評価することができる。本発明により得られたスコア行列は、タンパク質立体構造を予測するのに好適に使用される。

10

図面の簡単な説明

第1図は、本発明の一実施形態において使用されるハードウェア構成を示す図である。

- 15 第2図は、本発明のプロファイル行列間類似性評価システムを含む処理手順の一例を示すフローチャートである。

第3図は、本発明のプロファイル行列間類似性評価システムにおいて、各プロファイルカラムペア毎に類似性を評価し、スコア行列を作成するステップを示す図である。

- 20 第4図は、実施例1、比較例1及び比較例2において出力された予測結果の信頼度と感度とをプロットした図である。

第5図は、実施例1及び比較例3において出力された予測結果の信頼度と感度とをプロットした図である。

発明を実施するための最良の形態

- 25 以下、本発明を詳細に説明する。

1. 類似度評価システム

第1図は、本発明の一実施形態において使用されるハードウェア構成を示す図である。

第1図に示すように、本発明の類似性評価システムは、CPU101、ROM102、RAM103

、入力部104、情報通信送信/受信部105、出力部106、ハードディスクドライブ(HDD)107及びCD-ROMドライブ108等を備える。

CPU101は、情報記憶手段（例えば磁氣的及び／又は光学的記録媒体）に記憶されているプログラムに従って、類似性評価システム全体を制御する。そして、入力部104などから受け取った情報を出力部106に供給する。また、ネットワーク回線109を通じて受け取った情報に基づいて評価処理を実行することもできる。入力部104は、キーボードやマウス等であり、評価処理を実行する上で必要な条件又はデータを入力するときに操作される。ROM102は、本発明の類似性評価システムの動作に必要な処理を命令するプログラム等を格納する。RAM103は、類似性評価システムにおける処理を実行する上で必要なデータを一時的に格納する。

送信／受信部105は、CPU101の命令に基づいて、ネットワーク回線109等との間で情報通信（データの送受信処理）を実行するものであり、例えばモデム、ルーター等が例示される。出力部106は、入力手段104から入力されたプロファイルデータ、その他各種条件等を、CPU101からの命令に基づいて情報表示処理する（例えば表示画面、プリンタ）。CD-ROMドライブ108は、CPU101の指示に基づいて、CD-ROMに格納されている類似性評価システムを機能させるためのプログラム又はデータ等を読み出し、例えばRAM103に格納する。CD-ROMの代わりに記録媒体として書き換え可能なCD-R、CD-RWを用いることもできる。その場合には、CD-ROMドライブ108の代わりにCD-R又はCD-RW用ドライブを設ける。また、上記媒体の他に、DVD、MOとそれらの媒体を用い、それに対応するドライブを備える構成としてもよい。

コンピュータに本発明の類似性評価システムを機能させるためのプログラムは、例えばC言語等で書くことができる。従って、このソフトウェアはWindows（登録商標）95/98/2000、Linux（登録商標）、UNIX（登録商標）等の各種オペレーティングシステムで作動させることが可能である。

第2図は、本発明のプロファイル行列間類似性評価システムを含む処理手順の一例を示すフローチャートである。

第2図に示すように、本発明にかかる類似性評価システムでは、まず、比較したい2つのプロファイル行列（入力プロファイル行列と対象プロファイル行列）を用意し、続いてそれらの類似性を評価し、必要に応じて評価結果を出力する。

以下、各処理について詳細に説明する。

(a) プロファイル行列の用意 (S 1 0)

プロファイル行列を用意するステップでは、比較したい2つのプロファイル行列が用意(抽出)される(S 1 1、S 1 2)。ここで、2つのプロファイル行列のうち、一方(対象プロファイル行列)は、立体構造が既知である複数のタンパク質に基づいて作成されたプロファイル行列(第2図中、長さm)である。他方(入力プロファイル行列)は、立体構造を予測したいタンパク質(立体構造は未知であると既知であるとを問わない)を含む複数のタンパク質に基づいて作成されたプロファイル行列(第2図中、長さn)であることが好ましい。

プロファイル行列の作成方法としては、上述した従来知られている方法を採用することができ、特に制限はない。たとえば、ある一配列を入力値として、既存のプログラムであるPSI-BLASTを用いて、配列データベースに検索をかけてマルチプルアラインメントを作成し、このマルチプルアラインメントに基づいてプロファイル行列を作成してもよい。また、生物学的に相互に関連しあう複数のタンパク質のアミノ酸配列の一群を入力値として、既存のプログラムであるCLUSTALWを用いてマルチプルアラインメントを作成し、当該マルチプルアラインメントに基づいてプロファイル行列を作成してもよい。また、予め作成されたマルチプルアラインメントを入力値とし、このマルチプルアラインメントに基づいて作成してもよい。

ここで、プロファイル行列は、ある代表アミノ酸配列の全配列に基づいて作成されていてもよく、また、代表配列中のモチーフ領域等、一部の領域に基づいて作成されていてもよい。また、マルチプルアラインメントを作成する際に、経験的に導出されたギャップペナルティを導入してもよい。

また、必要に応じて、プロファイル行列として、アミノ酸種の出現頻度を、アミノ酸種のランダム出現頻度で割った行列(PSSM: Gribskov, M., et al., (1987) Proc. Natl. Acad. Sci. USA, 84, 4355-4358)を用いてもよい。

入力プロファイル行列は、たとえば、立体構造を予測したいタンパク質を代表アミノ酸配列として、この配列に基づいて作成することができる。また、対象プロファイル行列については、たとえば、SCOP (Murzin et al., J. Mol. Biol.

247(4):536-540 (1995))やCATH(Orengo et al., Structure 5(8):1093-1108 (1997))といったタンパク質構造分類データベースから取得したタンパク質のアミノ酸配列を代表配列とし、この配列に基づいて作成することができる。こうして得られた対象プロファイル行列は、代表配列ごとに予め作成しておき、対象プロファイル行列データベースとして保持しておくことが好ましい。

(b) 相関係数の算出 (プロファイル行列の比較評価) (S 20)

続いて、プロファイル行列の類似性評価ステップでは、上記のステップで用意した入力プロファイル行列の各プロファイルカラムと、対象プロファイル行列の各プロファイルカラムとの間の類似性を、各カラムペア毎に評価をする。

10 第3図は、各プロファイルカラムペア毎に類似性を評価し、スコア行列を作成するステップを模式的に示した図である。

本発明において、プロファイルカラム間の類似性は、プロファイルカラム間の相関係数を算出することによって行う。

たとえば、入力プロファイル行列を $X = x_1 x_2 \cdots x_p \cdots x_n$ (ただし、 x_p はアミノ酸残基位置 p におけるプロファイルカラム) とし、対象プロファイル行列を $Y = y_1 y_2 \cdots y_q \cdots y_m$ (ただし、 y_q はアミノ酸残基位置 q におけるプロファイルカラム) としたときに、プロファイルカラム x_p および y_q 間の類似性スコア $c_{q,p}$ は、下記の式によって与えられる。

【数2】

$$C_{pq} = \frac{\sum_a^j (x_{pa} - \bar{x}_p)(y_{qa} - \bar{y}_q)}{\sqrt{\sum_a^j (x_{pa} - \bar{x}_p)^2 \sum_a^j (y_{qa} - \bar{y}_q)^2}}$$

[式中、 x_{pa} = プロファイルカラム x_p の要素

y_{qa} = プロファイルカラム y_q の要素

\bar{x}_p = プロファイルカラム x_p の平均値

\bar{y}_q = プロファイルカラム y_q の平均値

j = プロファイルカラムの要素数 (通常 20) である。]

本発明では、プロファイルカラム間の類似性をプロファイルカラム間の相関係数によって評価する。このため、プロファイルカラム間の相関の程度によって、類似性スコアが +1 から -1 の値をとることになる。たとえば、2つのプロファイルカラム中の要素間に相関がある場合、即ちアミノ酸置換パターンの傾向に類似性が有る場合には、相関係数は +1 に近い数値を取ることになる。また、2つのプロファイルカラムの各要素が互いにランダムな値を取っている場合、即ちアミノ酸置換パターンの傾向に相関が無い場合、相関係数は 0 になり、アミノ酸置換パターンの傾向が全く反対の場合、相関係数は -1 になり、アミノ酸置換パターンの傾向性の類似-非類似を非常に自然な形で表現する事が出来る。

したがって、本発明では、アミノ酸残基の保存性が高い保存領域のような相関が高い領域では、高い類似性スコアが得られるため、保存領域の類似性を精度よく評価することができる。

また、本発明によれば、アミノ酸残基の保存性だけではなく、内積によって類似性を評価する従来の方法 (Rychlewski et alら) では不可能であった領域に関する類似性評価、たとえば、モチーフ内に存在する非保存位置や、タンパク質立体構造上露出していることが重要で極性のみが重大な意義を占める位置、あるいはその逆に埋没部分に位置し疎水性のみが保存されている位置といった、激しいアミノ酸置換があるもののその置換パターンに共通性があると考えられる領域に

についての類似性をより精度良く評価することが可能である。

例えば、あるzinc fingerモチーフを有する2つのプロファイル行列を比較した場合を考えたとする。そのモチーフは

C-[DES]-x-C-x(3)-I

- 5 と表記される。これは、1, 4, 8番目の残基にそれぞれC, C, Iの残基が保存されており、2番目の残基では、D又はE又はSが出現し、3番目および、5, 6, 7番目の残基では保存残基が特に無いことが表されている。内積によって類似性を評価する従来の方法では、この場合、1, 2, 4, 8番目の残基位置では、高い数値を与えるが、その他の位置では低い数値しか与えない。したがって、内積によって類似性を評価する従来の方法は、モチーフの一部については類似性を評価しているものの、モチーフ全体の類似性については精度よく評価していないということになる。

- 15 しかしながら、本発明によれば、1, 2, 4, 8番目の残基位置に高い数値を与えるだけでなく、3, 5, 6, 7番目の残基位置においても、保存残基が特に無いという置換パターンの類似性を評価することが可能で、これら残基位置でも高い数値を与える。したがって、本発明によれば、モチーフ全体としてのパターン情報の全てを評価することが可能となる。

- 20 なお、本発明における類似性評価システムは、モチーフ領域に限られず、立体構造を予測したいタンパク質の配列全体に適用することができる。すなわち、ギャップペナルティを導入して得られたプロファイル行列間の類似性評価にも、好適に適用することができる。

さらに、本発明によれば、スコア行列の各要素（類似性スコア）の平均値および標準偏差がほぼ一定値をとるため、類似性スコアに対する煩雑な正規化処理を施す必要がないというメリットもある。

- 25 (c) スコア行列の作成

プロファイルカラム間の相関係数（類似性スコア）は、各プロファイルカラムの全部又は一部の組合せについて算出され、これに基づいてスコア行列が作成される。スコア行列は、類似性スコアが各プロファイルカラムの全組合せについて算出された場合は、入力プロファイル行列の長さを行とし、対象プロファイル行

列の長さを列とする行列であり、類似性スコアが各プロファイルカラムの一部の組合せについて算出された場合は、その組合せの数に応じた行と列を持つ行列となる。

第2図の例では、類似性スコアは各プロファイルカラムの全組合せについて算出されており、入力プロファイル行列の長さが n 、対象プロファイル行列の長さが m であることから、類似性スコアは $m \times n$ 個生成される（S 2 2）。したがって、スコア行列は n 行 m 列となる。スコア行列は、比較したいプロファイル行列の長さ、及び算出される類似性スコアの数に応じた行列を予め定義し（S 2 1）、定義された行列の各カラムに、各プロファイルカラム間の相関係数を入力することにより作成することができる（S 2 3）。

本発明で得られたスコア行列によって、2つのプロファイル行列の最終スコア（行列間の類似性）を精度よく算出することができる。最終スコアは既知の手法により作成することができる。たとえば、第2図の例では、比較されるプロファイル行列のそれぞれの代表アミノ酸配列と、本発明によって得られたこれらのプロファイル行列間のスコア行列を入力値として、ダイナミックプログラミングを用いて最適パスを算出する（S 2 4）ことによって最終スコアを求めることができる（S 2 5）。

以上の操作を、対象プロファイル行列データベースに保持してある対象プロファイル行列のすべてに対して行うことが好ましい。

20 2. タンパク質立体構造の予測（S 3 0）

対象プロファイル行列ごとに得られた最終スコアは、タンパク質立体構造を予測するのに好適に使用される。たとえば、以下の既知の手順にしたがって処理をされる。

(1) 入力値

25 まず、予測対象配列を含む入力プロファイル行列と、立体構造が既知である代表アミノ酸配列を含む対象プロファイル行列との最終スコア、および各代表配列の長さが入力される。このとき、対象プロファイル行列データベース中に N 本の既知代表配列があれば、 N 個の最終スコアと配列長が入力されることになる。

(2) 最終スコアの長さ依存性の補正

予測対象配列を含む入力プロファイル行列と、各既知代表配列を含む対象プロファイル行列との最終スコアは、代表配列長に依存した関係が認められる為、次のような統計処理を行う。まず、X軸に各代表配列の長さの自然対数をとった値、Y軸に予測対象配列を含む入力プロファイル行列と各既知代表配列を含むプロファイル行列との最終スコアをプロットし、異常なはずれ値を除いて回帰直線を引く。各長さにおける最終スコアの平均値は回帰直線で表されるものとみなし、予測対象配列を含む入力プロファイル行列と各既知代表配列を含む対象プロファイル行列との最終スコアは、平均値からのずれで評価される。通常良く使用されるように、標準偏差を単位として、そのずれの度合いが測定される。

(3) ソート

平均値からのずれが（高得点側に）大きいもの程類似性が有るとみなされる。それ故、平均値からのずれが（高得点側に）大きい順にソートされ、予測構造の候補とされる。

(4) 予測構造としてのアラインメントとスコア出力

上でソートされた順に予測構造の候補として出力される。結果全てを出力するのは無意味なため、予測精度を考慮し経験的に求められた閾値以上の平均値からのずれを有する結果のみを出力する。この時、予測精度の指標として、標準偏差を単位として計算される平均値からのずれの度合いが表示される。

予測対象配列を含む入力プロファイル行列と、各既知代表配列を含む対象プロファイル行列とのアラインメントおよび最終スコアの結果は、ダイナミックプログラミングを用いて逐次計算された際のものを出力する。各既知代表配列は立体構造既知なので、このアラインメント出力が立体構造予測結果に相当する。

3. コンピュータプログラム

本発明は、コンピュータを、タンパク質の立体構造を予測するためのタンパク質プロファイル行列間の類似性を評価するシステムとして機能させるためのプログラムをも提供する。本発明のコンピュータプログラムは、以下の手段：

(a) 入力プロファイル行列と、対象プロファイル行列の2つのプロファイル行列を用意する手段と、

(b) 前記入カプロファイル行列の各プロファイルカラムと、前記対象プロファイ

ル行列の各プロファイルカラムとの間の相関係数を、各プロファイルカラムの全部又は一部の組合せについて算出する手段と、

(c) 前記相関係数からなるスコア行列を作成する手段とを含むものである。

- 5 本発明のプログラムには、上記必須の手段以外に、汎用のプログラムとして通常備えられる汎用手段を含んでもよい。そのような手段としては、各種データの格納手段、情報の送受信手段、ディスプレイ、プリンター等の表示・出力手段等を挙げることができる。

4. コンピュータ用記録媒体

- 10 本発明のプログラムは、コンピュータ読み取り可能な記録媒体又はコンピュータに接続しうる記憶手段に保存することができる。本発明のプログラムを含有するコンピュータ用記録媒体又は記憶手段も本発明に含まれる。記録媒体又は記憶手段としては、磁氣的媒体（フレキシブルディスク、ハードディスクなど）、光学的媒体（CD、DVDなど）、磁気光学的媒体（MO、MD）などが挙げられる。

15 【実施例】

以下、実施例により本発明をさらに具体的に説明する。但し、本発明はこれら実施例に限定されるものではない。

実施例 1

(1) 対象プロファイル行列データベースの構築

- 20 構造分類データベース S C O P (URL:<http://scop.mrc-lmb.cam.ac.uk/scop/>) release1.59 に基づく分類から、代表配列を取得した。その中から、単独ドメインを有し解像度2.5Å以内の構造データを有するタンパク質のアミノ酸配列948本を選択した。948本の代表配列各々に対してPSI-BLASTとアミノ酸配列データベース(NRDB:<ftp://ftp.ncbi.nlm.nih.gov>より取得)を用いて対象プロファイル行列を構築し、対象プロファイル行列データベースを完成させた。

25 ここで使用した「NRDB」には、現在知られているほぼ大部分のタンパク質アミノ酸配列が含まれている。PSI-BLASTを使うことで、このNRDBから各代表配列に生物学的に関連あると考えられる配列を自動的に収集し、さらにプロファイル行列も作成することが出来る。

(2) 入力プロファイル行列の作成

本発明にかかるシステムによって正しい構造予測がなされているかどうかを調べるため、予測対象配列として構造が既に知られている配列、すなわち、対象プロファイル行列を作成する際に使用した上記 9 4 8 本の代表配列を使用した。入力プロファイル行列は、これらの予測対象配列を順次使用して、対象プロファイル行列の場合と同様の操作、すなわち、PSI-BLASTとアミノ酸配列データベース(NRDB)を用いて構築した。

(3) 各プロファイル行列間の比較

続いて、上記で構築された予測対象配列（本実施例では 9 4 8 本の各代表配列）を含む入力プロファイル行列と、対象プロファイル行列データベース中の対象プロファイル行列との比較が順次なされた。この際、プロファイル行列間のスコア行列の各要素（類似性スコア）は、相関係数を用いて計算された。

こうして得られたプロファイル行列間のスコア行列を入力値として、ダイナミックプログラミングによってプロファイル行列間の最終スコアとアラインメントが出力された。

各入力プロファイル行列に対して、以上の操作を対象プロファイル行列データベースに構築されたすべての対象プロファイル行列について行った。

(4) 最終処理及び結果出力

評価の出力は、既に説明した方法に従って、9 4 8 予測について各々結果出力を行った。すなわち、入力プロファイル行列と対象プロファイル行列との各最終スコアおよび各代表配列の長さを入力し、最終スコアの長さ依存性の補正を行った。続いて、平均値からのずれが（高得点側に）大きい順にソートし、ソートされた順に予測構造の候補として出力した。

こうして出力された予測構造の候補と、既にわかっている正しい予測構造とを比較することにより、予測結果の信頼度と感度を算出し、この結果を第 4 図に示した。

比較例 1

実施例 1 で取得した 9 4 8 本の代表配列を用いて、配列類似性検索として一般的に用いられている P S I - B L A S T を用いて構造予測を行った。すなわち、

9 4 8 本の代表配列各々に対してPSI-BLASTとアミノ酸配列データベース(N R D B:ftp://ftp.ncbi.nlm.nih.govより取得)を用いて構築したプロファイル行列を入力値とし、9 4 8 本の代表配列に対して類似性検索を行い、予測構造の候補を出力した。

- 5 こうして出力された予測構造の候補と、既にわかっている正しい予測構造とを比較することにより、予測結果の信頼度と感度を算出し、この結果を第4図に示した。

比較例 2

- 実施例1で取得した9 4 8 本の代表配列を用いて、配列類似性検索として一般的に用いられている I M P A L A (Schaffer, A. A., Wolf, Y. I., Ponting, C. P., Koonin, E. V., Aravind, L., and Altschul, S. F. (1999) Bioinformatics. 015:1000-1011) を用いて構造予測を行った。すなわち、9 4 8 本の代表配列を入力値とし、9 4 8 本の代表配列各々に対して予め作成し構築したプロファイル行列データベース(実施例1で構築した対象プロファイル行列データベースを使用した)に対して類似性検索を行い、予測構造の候補を出力した。

こうして出力された予測構造の候補と、既にわかっている正しい予測構造とを比較することにより、予測結果の信頼度と感度を算出し、この結果を第4図に示した。

- 第4図から、比較例1および2の手法に比べて、信頼度0.98以降において、本発明にかかる実施例1が常に感度で勝っていることが示される。

比較例 3

プロファイル行列間のスコア行列の各要素(類似性スコア)を、内積法(Rychlewski et al. (2000), 9:p232-241)を用いて計算した以外は実施例1と同様の手法で予測構造の候補を出力した。

- 25 こうして出力された予測構造の候補と、既にわかっている正しい予測構造とを比較することにより、予測結果の信頼度と感度を算出し、この結果を第5図に示した。

実施例 2

(1) 対象プロファイル行列データベースの構築

配列は、構造分類データベース SCOP (URL: <http://scop.mrc-lmb.cam.ac.uk/scop/>) release1.59に基づく分類から、お互いの同一残基率が40%未満であるドメイン単位の代表配列4381本を、SCOPの配列データベースであるASTRAL (<http://astral.stanford.edu/>) データベースから取得した。更に、タンパク質立体構造データベース PDB (URL: <http://www.rcsb.org/pdb/>) に登録されているが、SCOPに未登録であるものであって、ASTRALから取得した上記4381本の配列と非類似のものを下記 (A) ~ (D) の要領で取得し、代表配列に加えた。このようにして選択されたアミノ酸配列各々に対して、下記 (A) ~ (D) の要領で PSI-BLAST と N R D B を用いて対象プロファイル行列を構築し、対象プロファイル行列データベースを完成させた。

(A) 対象プロファイル行列データベース A の構築

2002年5月18日時点での PDB 中のアミノ酸配列を SCOP release1.59 の分類に基づく代表配列に対して BLASTP (Altschul et al., Nucleic Acids Res. (1997) 25(17): 3389-3402: 非特許文献 2) をかけ、期待値が 0.00001 以上のものを選んだ。さらにそれらを配列のクラスタリングを行うプログラムである blastclust 15 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100 105 110 115 120 125 130 135 140 145 150 155 160 165 170 175 180 185 190 195 200 205 210 215 220 225 230 235 240 245 250 255 260 265 270 275 280 285 290 295 300 305 310 315 320 325 330 335 340 345 350 355 360 365 370 375 380 385 390 395 400 405 410 415 420 425 430 435 440 445 450 455 460 465 470 475 480 485 490 495 500 505 510 515 520 525 530 535 540 545 550 555 560 565 570 575 580 585 590 595 600 605 610 615 620 625 630 635 640 645 650 655 660 665 670 675 680 685 690 695 700 705 710 715 720 725 730 735 740 745 750 755 760 765 770 775 780 785 790 795 800 805 810 815 820 825 830 835 840 845 850 855 860 865 870 875 880 885 890 895 900 905 910 915 920 925 930 935 940 945 950 955 960 965 970 975 980 985 990 995 1000 1005 1010 1015 1020 1025 1030 1035 1040 1045 1050 1055 1060 1065 1070 1075 1080 1085 1090 1095 1100 1105 1110 1115 1120 1125 1130 1135 1140 1145 1150 1155 1160 1165 1170 1175 1180 1185 1190 1195 1200 1205 1210 1215 1220 1225 1230 1235 1240 1245 1250 1255 1260 1265 1270 1275 1280 1285 1290 1295 1300 1305 1310 1315 1320 1325 1330 1335 1340 1345 1350 1355 1360 1365 1370 1375 1380 1385 1390 1395 1400 1405 1410 1415 1420 1425 1430 1435 1440 1445 1450 1455 1460 1465 1470 1475 1480 1485 1490 1495 1500 1505 1510 1515 1520 1525 1530 1535 1540 1545 1550 1555 1560 1565 1570 1575 1580 1585 1590 1595 1600 1605 1610 1615 1620 1625 1630 1635 1640 1645 1650 1655 1660 1665 1670 1675 1680 1685 1690 1695 1700 1705 1710 1715 1720 1725 1730 1735 1740 1745 1750 1755 1760 1765 1770 1775 1780 1785 1790 1795 1800 1805 1810 1815 1820 1825 1830 1835 1840 1845 1850 1855 1860 1865 1870 1875 1880 1885 1890 1895 1900 1905 1910 1915 1920 1925 1930 1935 1940 1945 1950 1955 1960 1965 1970 1975 1980 1985 1990 1995 2000 2005 2010 2015 2020 2025 2030 2035 2040 2045 2050 2055 2060 2065 2070 2075 2080 2085 2090 2095 2100 2105 2110 2115 2120 2125 2130 2135 2140 2145 2150 2155 2160 2165 2170 2175 2180 2185 2190 2195 2200 2205 2210 2215 2220 2225 2230 2235 2240 2245 2250 2255 2260 2265 2270 2275 2280 2285 2290 2295 2300 2305 2310 2315 2320 2325 2330 2335 2340 2345 2350 2355 2360 2365 2370 2375 2380 2385 2390 2395 2400 2405 2410 2415 2420 2425 2430 2435 2440 2445 2450 2455 2460 2465 2470 2475 2480 2485 2490 2495 2500 2505 2510 2515 2520 2525 2530 2535 2540 2545 2550 2555 2560 2565 2570 2575 2580 2585 2590 2595 2600 2605 2610 2615 2620 2625 2630 2635 2640 2645 2650 2655 2660 2665 2670 2675 2680 2685 2690 2695 2700 2705 2710 2715 2720 2725 2730 2735 2740 2745 2750 2755 2760 2765 2770 2775 2780 2785 2790 2795 2800 2805 2810 2815 2820 2825 2830 2835 2840 2845 2850 2855 2860 2865 2870 2875 2880 2885 2890 2895 2900 2905 2910 2915 2920 2925 2930 2935 2940 2945 2950 2955 2960 2965 2970 2975 2980 2985 2990 2995 3000 3005 3010 3015 3020 3025 3030 3035 3040 3045 3050 3055 3060 3065 3070 3075 3080 3085 3090 3095 3100 3105 3110 3115 3120 3125 3130 3135 3140 3145 3150 3155 3160 3165 3170 3175 3180 3185 3190 3195 3200 3205 3210 3215 3220 3225 3230 3235 3240 3245 3250 3255 3260 3265 3270 3275 3280 3285 3290 3295 3300 3305 3310 3315 3320 3325 3330 3335 3340 3345 3350 3355 3360 3365 3370 3375 3380 3385 3390 3395 3400 3405 3410 3415 3420 3425 3430 3435 3440 3445 3450 3455 3460 3465 3470 3475 3480 3485 3490 3495 3500 3505 3510 3515 3520 3525 3530 3535 3540 3545 3550 3555 3560 3565 3570 3575 3580 3585 3590 3595 3600 3605 3610 3615 3620 3625 3630 3635 3640 3645 3650 3655 3660 3665 3670 3675 3680 3685 3690 3695 3700 3705 3710 3715 3720 3725 3730 3735 3740 3745 3750 3755 3760 3765 3770 3775 3780 3785 3790 3795 3800 3805 3810 3815 3820 3825 3830 3835 3840 3845 3850 3855 3860 3865 3870 3875 3880 3885 3890 3895 3900 3905 3910 3915 3920 3925 3930 3935 3940 3945 3950 3955 3960 3965 3970 3975 3980 3985 3990 3995 4000 4005 4010 4015 4020 4025 4030 4035 4040 4045 4050 4055 4060 4065 4070 4075 4080 4085 4090 4095 4100 4105 4110 4115 4120 4125 4130 4135 4140 4145 4150 4155 4160 4165 4170 4175 4180 4185 4190 4195 4200 4205 4210 4215 4220 4225 4230 4235 4240 4245 4250 4255 4260 4265 4270 4275 4280 4285 4290 4295 4300 4305 4310 4315 4320 4325 4330 4335 4340 4345 4350 4355 4360 4365 4370 4375 4380 4385 4390 4395 4400 4405 4410 4415 4420 4425 4430 4435 4440 4445 4450 4455 4460 4465 4470 4475 4480 4485 4490 4495 4500 4505 4510 4515 4520 4525 4530 4535 4540 4545 4550 4555 4560 4565 4570 4575 4580 4585 4590 4595 4600 4605 4610 4615 4620 4625 4630 4635 4640 4645 4650 4655 4660 4665 4670 4675 4680 4685 4690 4695 4700 4705 4710 4715 4720 4725 4730 4735 4740 4745 4750 4755 4760 4765 4770 4775 4780 4785 4790 4795 4800 4805 4810 4815 4820 4825 4830 4835 4840 4845 4850 4855 4860 4865 4870 4875 4880 4885 4890 4895 4900 4905 4910 4915 4920 4925 4930 4935 4940 4945 4950 4955 4960 4965 4970 4975 4980 4985 4990 4995 5000 5005 5010 5015 5020 5025 5030 5035 5040 5045 5050 5055 5060 5065 5070 5075 5080 5085 5090 5095 5100 5105 5110 5115 5120 5125 5130 5135 5140 5145 5150 5155 5160 5165 5170 5175 5180 5185 5190 5195 5200 5205 5210 5215 5220 5225 5230 5235 5240 5245 5250 5255 5260 5265 5270 5275 5280 5285 5290 5295 5300 5305 5310 5315 5320 5325 5330 5335 5340 5345 5350 5355 5360 5365 5370 5375 5380 5385 5390 5395 5400 5405 5410 5415 5420 5425 5430 5435 5440 5445 5450 5455 5460 5465 5470 5475 5480 5485 5490 5495 5500 5505 5510 5515 5520 5525 5530 5535 5540 5545 5550 5555 5560 5565 5570 5575 5580 5585 5590 5595 5600 5605 5610 5615 5620 5625 5630 5635 5640 5645 5650 5655 5660 5665 5670 5675 5680 5685 5690 5695 5700 5705 5710 5715 5720 5725 5730 5735 5740 5745 5750 5755 5760 5765 5770 5775 5780 5785 5790 5795 5800 5805 5810 5815 5820 5825 5830 5835 5840 5845 5850 5855 5860 5865 5870 5875 5880 5885 5890 5895 5900 5905 5910 5915 5920 5925 5930 5935 5940 5945 5950 5955 5960 5965 5970 5975 5980 5985 5990 5995 6000 6005 6010 6015 6020 6025 6030 6035 6040 6045 6050 6055 6060 6065 6070 6075 6080 6085 6090 6095 6100 6105 6110 6115 6120 6125 6130 6135 6140 6145 6150 6155 6160 6165 6170 6175 6180 6185 6190 6195 6200 6205 6210 6215 6220 6225 6230 6235 6240 6245 6250 6255 6260 6265 6270 6275 6280 6285 6290 6295 6300 6305 6310 6315 6320 6325 6330 6335 6340 6345 6350 6355 6360 6365 6370 6375 6380 6385 6390 6395 6400 6405 6410 6415 6420 6425 6430 6435 6440 6445 6450 6455 6460 6465 6470 6475 6480 6485 6490 6495 6500 6505 6510 6515 6520 6525 6530 6535 6540 6545 6550 6555 6560 6565 6570 6575 6580 6585 6590 6595 6600 6605 6610 6615 6620 6625 6630 6635 6640 6645 6650 6655 6660 6665 6670 6675 6680 6685 6690 6695 6700 6705 6710 6715 6720 6725 6730 6735 6740 6745 6750 6755 6760 6765 6770 6775 6780 6785 6790 6795 6800 6805 6810 6815 6820 6825 6830 6835 6840 6845 6850 6855 6860 6865 6870 6875 6880 6885 6890 6895 6900 6905 6910 6915 6920 6925 6930 6935 6940 6945 6950 6955 6960 6965 6970 6975 6980 6985 6990 6995 7000 7005 7010 7015 7020 7025 7030 7035 7040 7045 7050 7055 7060 7065 7070 7075 7080 7085 7090 7095 7100 7105 7110 7115 7120 7125 7130 7135 7140 7145 7150 7155 7160 7165 7170 7175 7180 7185 7190 7195 7200 7205 7210 7215 7220 7225 7230 7235 7240 7245 7250 7255 7260 7265 7270 7275 7280 7285 7290 7295 7300 7305 7310 7315 7320 7325 7330 7335 7340 7345 7350 7355 7360 7365 7370 7375 7380 7385 7390 7395 7400 7405 7410 7415 7420 7425 7430 7435 7440 7445 7450 7455 7460 7465 7470 7475 7480 7485 7490 7495 7500 7505 7510 7515 7520 7525 7530 7535 7540 7545 7550 7555 7560 7565 7570 7575 7580 7585 7590 7595 7600 7605 7610 7615 7620 7625 7630 7635 7640 7645 7650 7655 7660 7665 7670 7675 7680 7685 7690 7695 7700 7705 7710 7715 7720 7725 7730 7735 7740 7745 7750 7755 7760 7765 7770 7775 7780 7785 7790 7795 7800 7805 7810 7815 7820 7825 7830 7835 7840 7845 7850 7855 7860 7865 7870 7875 7880 7885 7890 7895 7900 7905 7910 7915 7920 7925 7930 7935 7940 7945 7950 7955 7960 7965 7970 7975 7980 7985 7990 7995 8000 8005 8010 8015 8020 8025 8030 8035 8040 8045 8050 8055 8060 8065 8070 8075 8080 8085 8090 8095 8100 8105 8110 8115 8120 8125 8130 8135 8140 8145 8150 8155 8160 8165 8170 8175 8180 8185 8190 8195 8200 8205 8210 8215 8220 8225 8230 8235 8240 8245 8250 8255 8260 8265 8270 8275 8280 8285 8290 8295 8300 8305 8310 8315 8320 8325 8330 8335 8340 8345 8350 8355 8360 8365 8370 8375 8380 8385 8390 8395 8400 8405 8410 8415 8420 8425 8430 8435 8440 8445 8450 8455 8460 8465 8470 8475 8480 8485 8490 8495 8500 8505 8510 8515 8520 8525 8530 8535 8540 8545 8550 8555 8560 8565 8570 8575 8580 8585 8590 8595 8600 8605 8610 8615 8620 8625 8630 8635 8640 8645 8650 8655 8660 8665 8670 8675 8680 8685 8690 8695 8700 8705 8710 8715 8720 8725 8730 8735 8740 8745 8750 8755 8760 8765 8770 8775 8780 8785 8790 8795 8800 8805 8810 8815 8820 8825 8830 8835 8840 8845 8850 8855 8860 8865 8870 8875 8880 8885 8890 8895 8900 8905 8910 8915 8920 8925 8930 8935 8940 8945 8950 8955 8960 8965 8970 8975 8980 8985 8990 8995 9000 9005 9010 9015 9020 9025 9030 9035 9040 9045 9050 9055 9060 9065 9070 9075 9080 9085 9090 9095 9100 9105 9110 9115 9120 9125 9130 9135 9140 9145 9150 9155 9160 9165 9170 9175 9180 9185 9190 9195 9200 9205 9210 9215 9220 9225 9230 9235 9240 9245 9250 9255 9260 9265 9270 9275 9280 9285 9290 9295 9300 9305 9310 9315 9320 9325 9330 9335 9340 9345 9350 9355 9360 9365 9370 9375 9380 9385 9390 9395 9400 9405 9410 9415 9420 9425 9430 9435 9440 9445 9450 9455 9460 9465 9470 9475 9480 9485 9490 9495 9500 9505 9510 9515 9520 9525 9530 9535 9540 9545 9550 9555 9560 9565 9570 9575 9580 9585 9590 9595 9600 9605 9610 9615 9620 9625 9630 9635 9640 9645 9650 9655 9660 9665 9670 9675 9680 9685 9690 9695 9700 9705 9710 9715 9720 9725 9730 9735 9740 9745 9750 9755 9760 9765 9770 9775 9780 9785 9790 9795 9800 9805 9810 9815 9820 9825 9830 9835 9840 9845 9850 9855 9860 9865 9870 9875 9880 9885 9890 9895 9900 9905 9910 9915 9920 9925 9930 9935 9940 9945 9950 9955 9960 9965 9970 9975 9980 9985 9990 9995 10000 10005 10010 10015 10020 10025 10030 10035 10040 10045 10050 10055 10060 10065 10070 10075 10080 10085 10090 10095 10100 10105 10110 10115 10120 10125 10130 10135 10140 10145 10150 10155 10160 10165 10170 10175 10180 10185 10190 10195 10200 10205 10210 10215 10220 10225 10230 10235 10240 10245 10250 10255 10260 10265 10270 10275 10280 10285 10290 10295 10300 10305 10310 10315 10320 10325 10330 10335 10340 10345 10350 10355 10360 10365 10370 10375 10380 10385 10390 10395 10400 10405 10410 10415 10420 10425 10430 10435 10440 10445 10450 10455 10460 10465 10470 10475 10480 10485 10490 10495 10500 10505 10510 10515 10520 10525 10530 10535 10540 10545 10550 10555 10560 10565 10570 10575 10580 10585 10590 10595 10600 10605 10610 10615 10620 10625 10630 10635 10640 10645 10650 10655 10660 10665 10670 10675 10680 10685 10690 10695 10700 10705 10710 10715 10720 10725 10730 10735 10740 10745 10750 10755 10760 10765 10770 10775 10780 10785 10790 10795 10800 10805 10810 10815 10820 10825 10830 10835 10840 10845 10850 10855 10860 10865 10870 10875 10880 10885 10890 10895 10900 10905 10910 10915 10920 10925 10930 10935 10940 10945 10950 10955 10960 10965 10970 10975 10980 10985 10990 10995 11000 11005 11010 11015 11020 11025 11030 11035 11040 11045 11050 11055 11060 11065 11070 11075 11080 11085 11090 11095 11100 11105 11110 11115 11120 11125 11130 11135 11140 11145 11150 11155 11160 11165 11170 11175 11180 11185 11190 11195 11200 11205 11210 11215 11220 11225 11230 11235 11240 11245 11250 11255 11260 11265 11270 11275 11280 11285 11290 11295 11300 11305 11310 11315 11320 11325 11330 11335 11340 11345 11350 11355 11360 11365 11370 11375 11380 11385 11390 11395 11400 11405 11410 11415 11420 11425 11430 11435 11440 11445 11450 11455 11460 11465 11470 11475 11480 11485 11490 11495 11500 11505 11510 11515 11520 11525 11530 11535 11540 11545 11550 11555 11560 11565 11570 11575 11580 11585 11590 11595 11600 11605 11610 11615 11620 11625 11630 11635 11640 11645 11650 11655 11660 11665 11670 11675 11680 11685 116

2002 年 7 月 14 日時点での PDB と 2002 年 6 月 23 日時点での PDB 中のアミノ酸配列の差分を上記 (B) で作成した代表配列に対して BLASTP をかけ、期待値が 0.00001 以上のものを選んだ。さらにそれらを blastclust にかけて、互いの同一残基率が 40%未満となるように配列 23 本を選択した。このようにして選択された配列と、上記 (B) で作成した代表配列との合計 4701 本の配列各々に対して、PSI-BLAST と 2002 年 7 月 9 日時点の NRDB を用いて対象プロファイル行列を構築し、対象プロファイル行列データベース C を完成させた。

(D) 対象プロファイル行列データベース D の構築

上記 (C) で作成した代表配列の合計 4701 本の配列各々に対して、PSI-BLAST と 2002 年 8 月 6 日時点の NRDB を用いて対象プロファイル行列を構築し、対象プロファイル行列データベース D を完成させた。

(2) 入力プロファイル行列の作成

配列は、隔年で行われる世界的規模で行われる構造予測コンテストの 2002 年度大会である CASP5/CAFASP3 (URL: [http:// predictioncenter.llnl.gov/casp5/](http://predictioncenter.llnl.gov/casp5/)) において、構造認識部門 (通常の配列解析手法では立体構造既知であるタンパク質と明白な配列類似性を有さないが、その構造が (実際に解かれてみると) 既知立体構造との構造類似性を有する、即ち類似性検索が困難なタンパク質に関する予測部門) において出題された配列、すなわち、現在通常の配列解析手法 (例えば、PSI-BLAST など) では、立体構造既知であるタンパク質と明白な配列類似性を有さないタンパク質であり、かつ、その構造が (実際に解かれてみると) 既知立体構造との構造類似性が明らかになったアミノ酸配列を用いた。具体的には、URL: <http://www.cs.bgu.ac.il/~dfischer/CAFASP3/targets.html> において、下記のターゲット番号が付されたアミノ酸配列 22 本を用いた。

T0130、T0132、T0134、T0135、T0136、T0138、T0146、T0147、T0148、T0156、T0157、T0159、T0162、T0168、T0170、T0172、T0173、T0174、T0186、T0187、T0191、T0193

これら 22 本の配列各々に対して、PSI-BLAST と NRDB を用いて入力プロファイル行列を構築し、入力プロファイル行列データベースを完成させた。

なお、NRDB としては、2002 年 5 月 18 日時点、2002 年 6 月 17 日時点、2002 年 7 月 9 日時点、及び 2002 年 8 月 6 日時点のものの計 4 種類を使用し、得られた

入力プロファイル行列データベースを、それぞれ、「入力プロファイル行列データベースA」、「入力プロファイル行列データベースB」、「入力プロファイル行列データベースC」、及び「入力プロファイル行列データベースD」とした。

(3) 各プロファイル行列間の比較

- 5 続いて、上記で構築された予測対象配列を含む入力プロファイル行列データベースAの入力プロファイル行列と、対象プロファイル行列データベースA中の対象プロファイル行列との比較を、実施例1の「(3)各プロファイル行列間の比較」と同様の手順で行った（比較A）。

- 10 同様の操作を、入力プロファイル行列データベースBと対象プロファイル行列データベースBに対して、入力プロファイル行列データベースCと対象プロファイル行列データベースCに対して、及び、入力プロファイル行列データベースDと対象プロファイル行列データベースDに対して、それぞれ行った（比較B, C, D）。

(4) 最終処理及び結果出力

- 15 評価の出力は、既に説明した方法に従って22予測について各々結果出力を行った。即ち、各データベースの組合せ（比較A～D）においてそれぞれ得られた、入力プロファイル行列と対象プロファイル行列との各最終スコアおよび、各代表配列の長さを入力し、最終スコアの長さ依存性を補正した。続いて平均値からのずれが、（高得点側に）大きい順にソートし、ソートされた順に上位10個までを
20 予測構造の候補として22本の配列各々に対して出力した（出力A～D）。

- こうして出力された予測構造の候補と、コンテストの予測構造投稿期間の後に公開された実験により解かれた立体構造とを比較することで、予測結果の正確さが測定された。予測構造評価方法の一つは、予測構造と正解構造の重ね合わせを行い、対応残基が3Åより短い距離にある残基数を出力A～Dについて積算すること（sum値）により行われた。22のタンパク質を構造ドメイン単位（全部で34ドメイン）で眺めた結果によれば、構造予測コンテストCASP5/CAFASP3における上記構造認識部門において22本の配列各々に対して上位1個の予測を考慮した時、本手法のsum値は「577」であり、これは、配列情報を用いた他のいかなる手法よりも優れているものであった。
- 25

また、ある閾値を設定してある入力（予測対象）配列に対する予測の成否を観測した場合でも、22本の配列各々に対して上位1個の予測を考慮した時本手法は、予測が成功したと判断される個数を出力A～Dについて積算したもの（correct 値）において、「9」と高く、配列情報を用いた他のいかなる手法よりも優れて

5 いることが示された。

請求の範囲

1. タンパク質の立体構造を予測するためのタンパク質プロファイル行列間の類似性を評価するシステムであって、

前記プロファイル行列は、関連する複数のタンパク質のアミノ酸配列を多重並置させたマルチプルアラインメントにおいて、各アミノ酸残基位置におけるアミノ酸種毎の出現確率を備えたプロファイルカラムの群から構成され、

前記類似性評価システムは、以下の手段：

(a) 立体構造を予測したいタンパク質を含む複数のタンパク質に基づいて作成される入力プロファイル行列と、立体構造が既知である複数のタンパク質に基づいて作成される対象プロファイル行列の2つのプロファイル行列を用意する手段と、

(b) 前記入力プロファイル行列の各プロファイルカラムと、前記対象プロファイル行列の各プロファイルカラムとの間の相関係数を、各プロファイルカラムの全部又は一部の組合せについて算出する手段と、

(c) 前記相関係数からなるスコア行列を作成する手段とを含むシステム。

2. 請求の範囲第1項に記載のシステムにより作成されたスコア行列を用いることを特徴とするタンパク質立体構造の予測システム。

3. コンピュータを、タンパク質の立体構造を予測するためのタンパク質プロファイル行列間の類似性を評価するシステムとして機能させるためのプログラムであって、

前記プロファイル行列は、関連する複数のタンパク質のアミノ酸配列を多重並置させたマルチプルアラインメントにおいて、各アミノ酸残基位置におけるアミノ酸種毎の出現確率を備えたプロファイルカラムの群から構成され、

前記類似性評価システムは、以下の手段：

(a) 立体構造を予測したいタンパク質を含む複数のタンパク質に基づいて作成される入力プロファイル行列と、立体構造が既知である複数のタンパク質に基づいて作成される対象プロファイル行列の2つのプロファイル行列を用意する手段と、

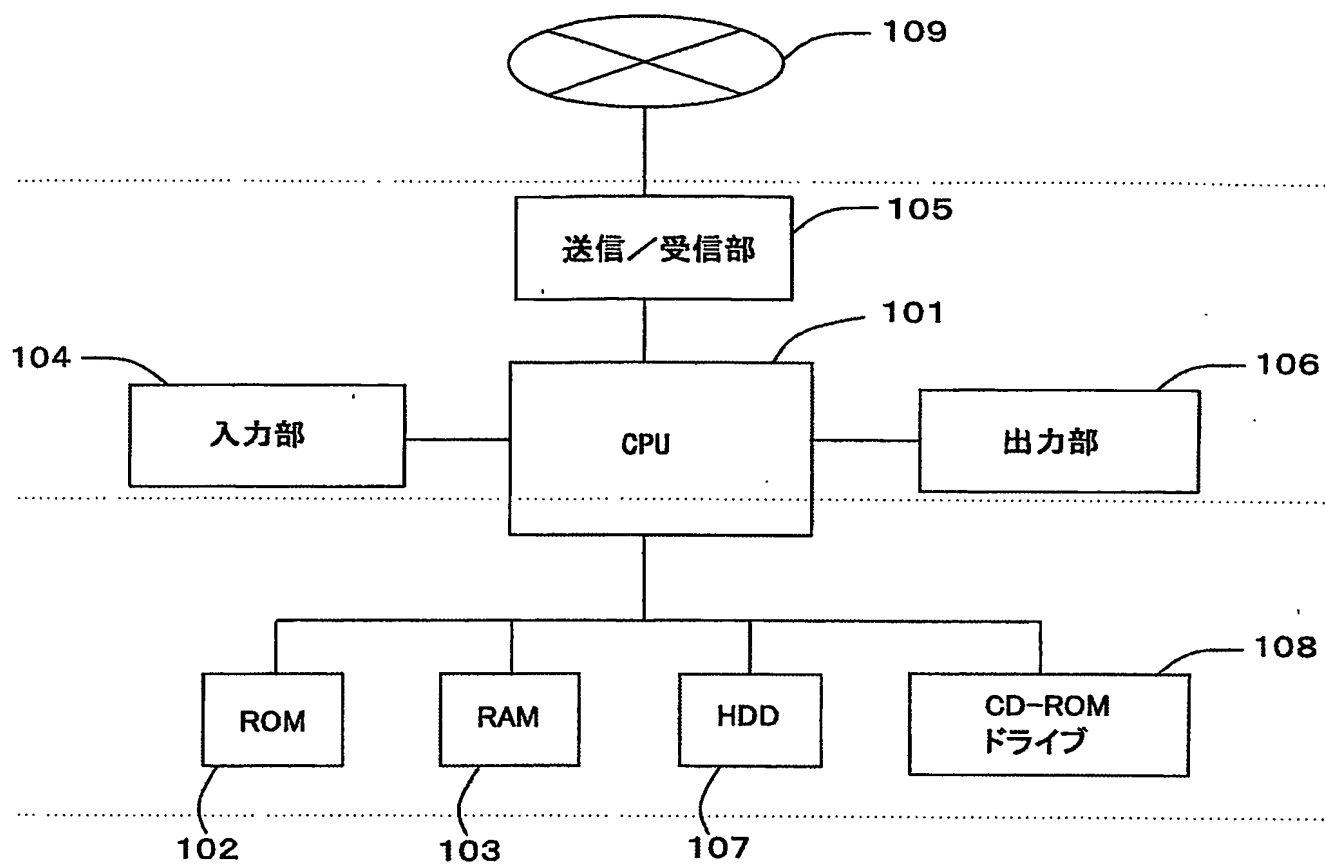
(b) 前記入カプロファイル行列の各プロファイルカラムと、前記対象プロファイル行列の各プロファイルカラムとの間の相関係数を、各プロファイルカラムの全部又は一部の組合せについて算出する手段と、

(c) 前記相関係数からなるスコア行列を作成する手段と

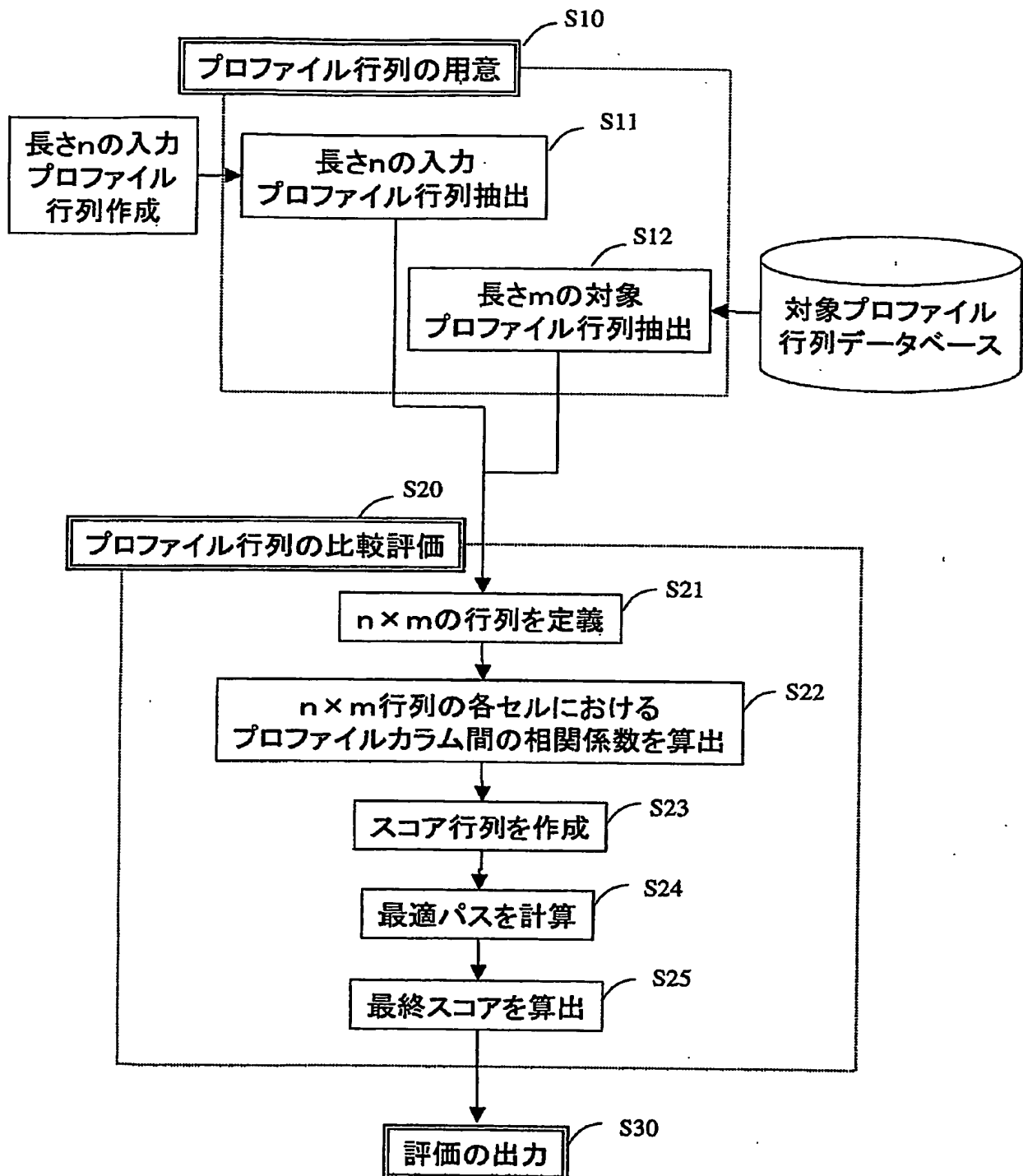
5 を含むプログラム。

4. 請求の範囲第3項に記載のプログラムを記録したコンピュータ読み取り可能な記録媒体。

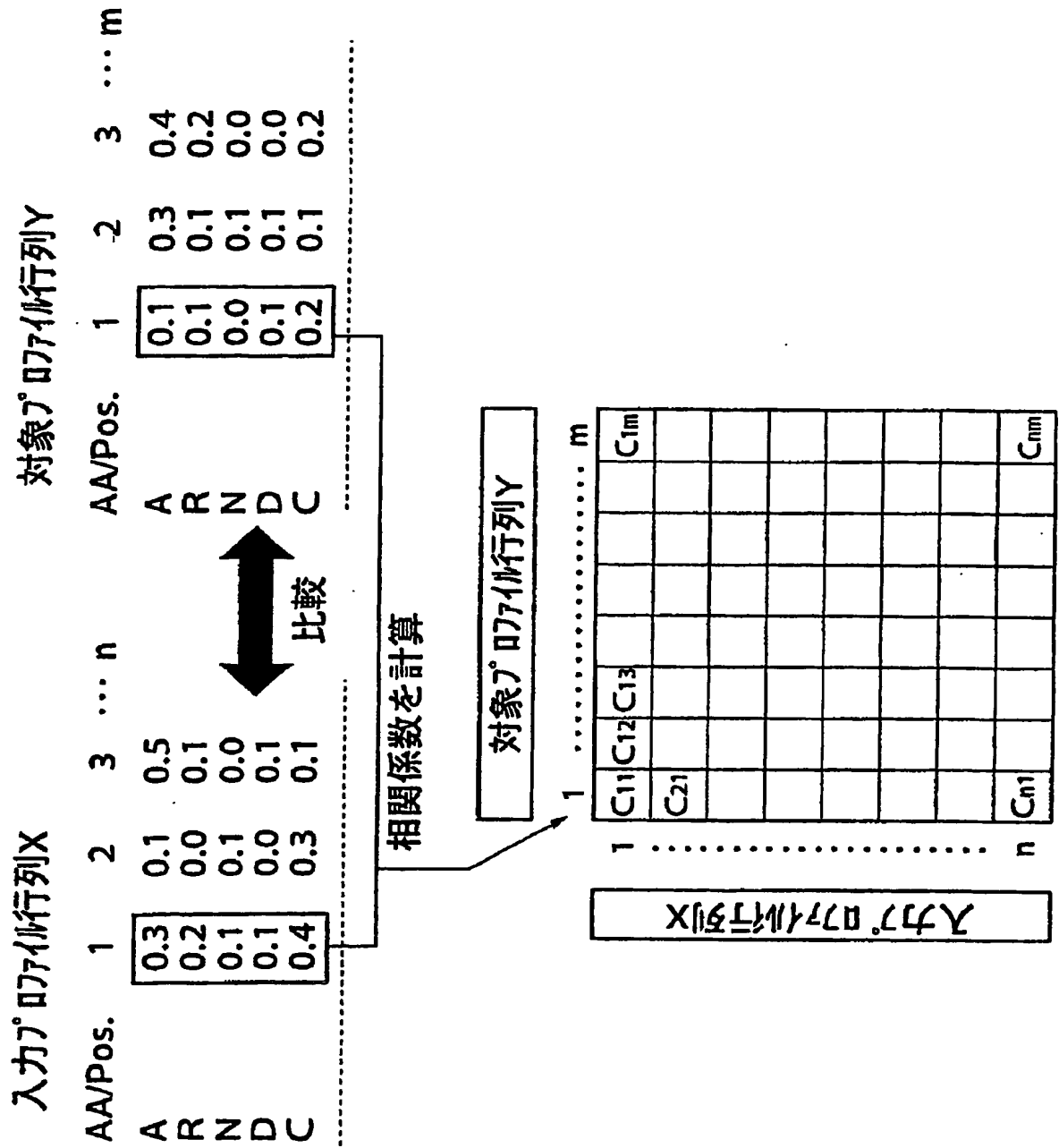
第 1 図



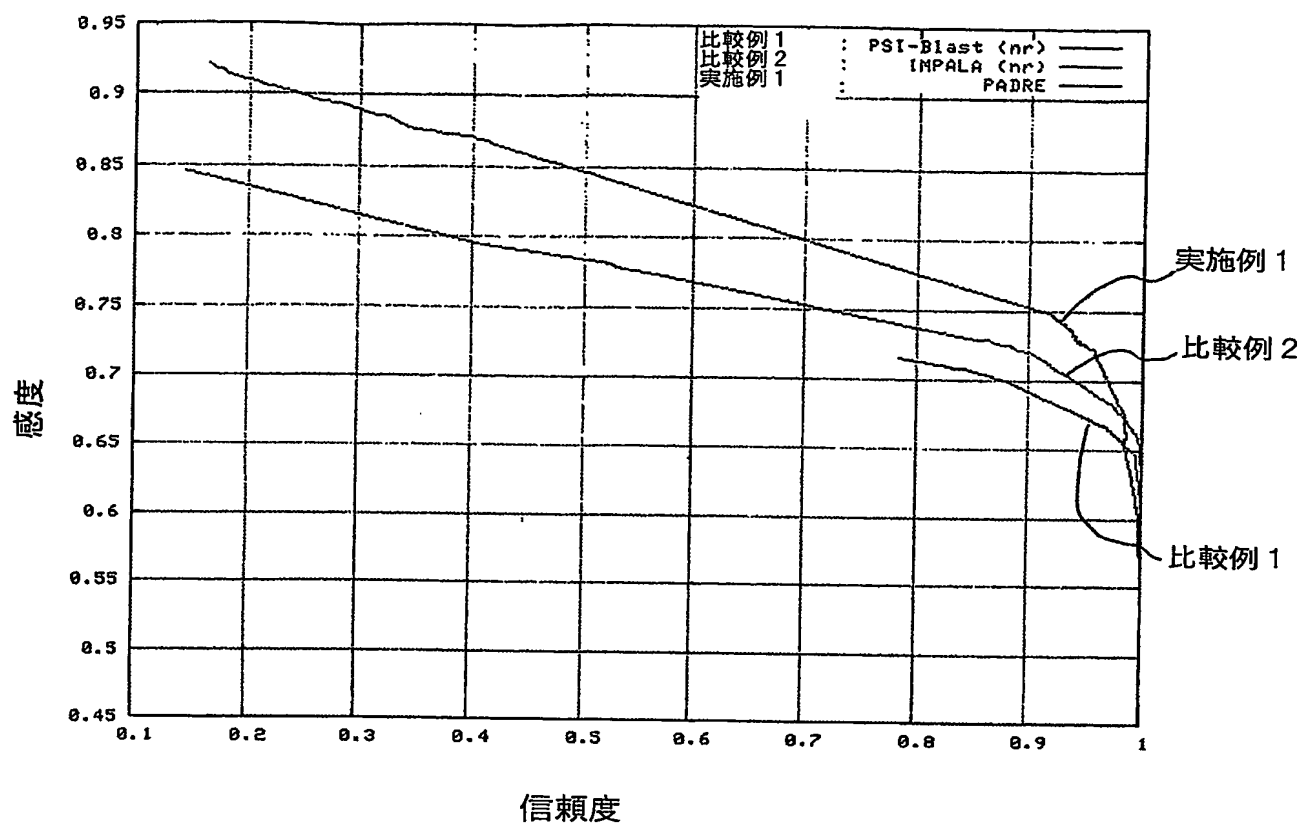
第 2 図



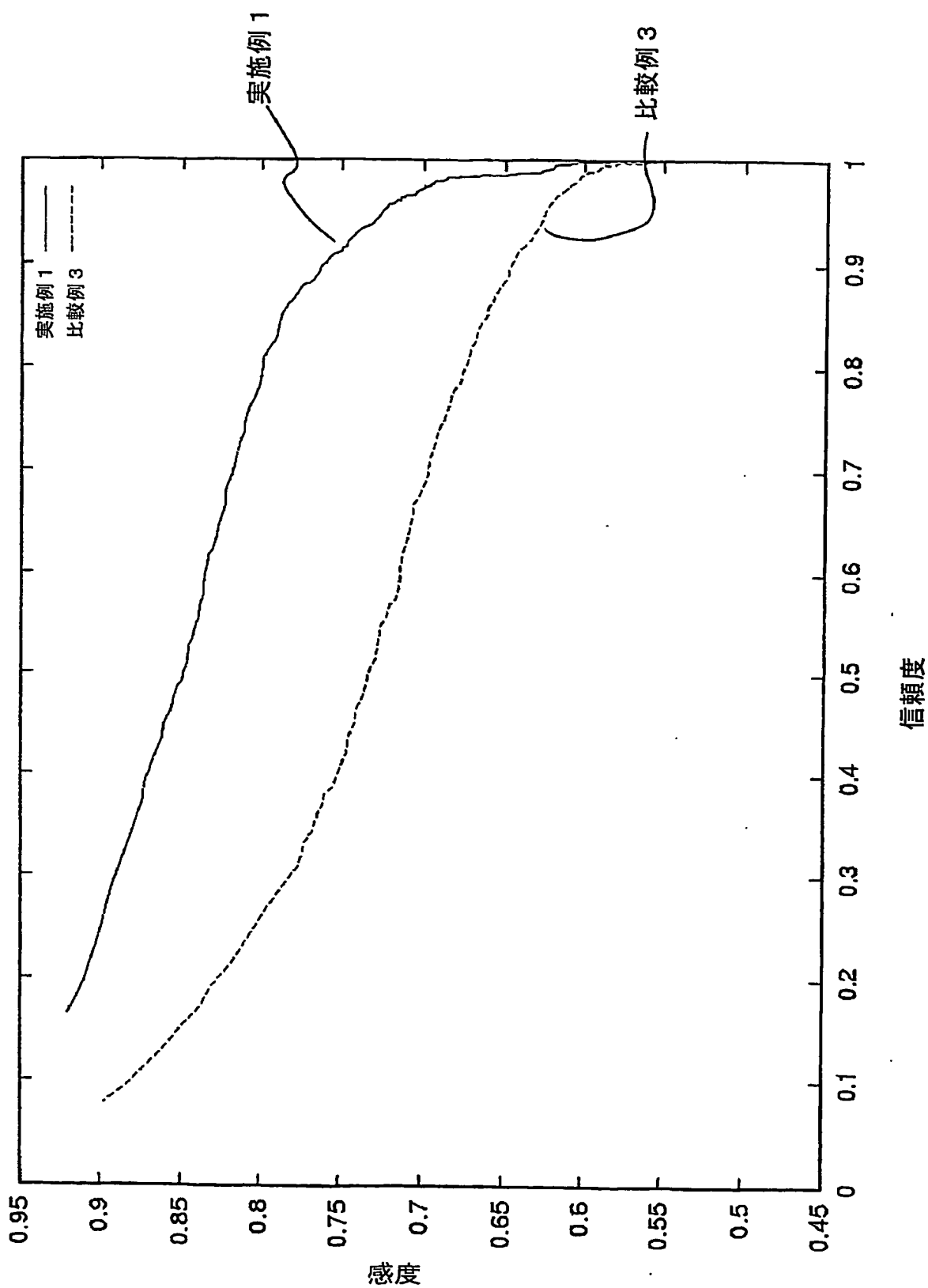
第3図



第4図



第 5 図



INTERNATIONAL SEARCH REPORT

International application No.

PCT/JP03/16982

A. CLASSIFICATION OF SUBJECT MATTER
Int.Cl⁷ G06F19/00

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
Int.Cl⁷ G06F19/00

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched
Jitsuyo Shinan Koho 1922-1996 Toroku Jitsuyo Shinan Koho 1994-2004
Kokai Jitsuyo Shinan Koho 1971-2004 Jitsuyo Shinan Toroku Koho 1996-2004

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
JICST FILE (JOIS), WPI, INSPEC (DIALOG)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	YONA, G. LEVITT, M. 'Within the Twilight Zone: A Sensitive Profile-Profile Comparison Tool Based Information Theory', Journal of Molecular Biology, 01 February, 2002 (01.02.02), Vol.315, Issue 5, pages 1257 to 1275; especially, pages 1259 to 1260 [on line] [retrieved on 29 January, 2004 (29.01.04)], Retrieved from: <URL=http://www.cs.cornell.edu/golan/Papers/jmb02.pdf>	1-4
A	RYCHLEWSKI, L. et al., 'Comparison of sequence profiles. Strategies for structural predictions using sequence information', Protein Science, February 2000, Vol.9, Issue 2, pages 232 to 241 [on line] [retrieved on 29 January, 2004 (29.01.04)], retrieved from <http://www.proteinscience.org/cgi/reprint/9/2/232.pdf>	1-4

☒ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

<p>* Special categories of cited documents:</p> <p>"A" document defining the general state of the art which is not considered to be of particular relevance</p> <p>"E" earlier document but published on or after the international filing date</p> <p>"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)</p> <p>"O" document referring to an oral disclosure, use, exhibition or other means</p> <p>"P" document published prior to the international filing date but later than the priority date claimed</p>	<p>"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention</p> <p>"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone</p> <p>"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art</p> <p>"&" document member of the same patent family</p>
--	---

Date of the actual completion of the international search
30 January, 2004 (30.01.04)

Date of mailing of the international search report
10 February, 2004 (10.02.04)

Name and mailing address of the ISA/
Japanese Patent Office

Authorized officer

Facsimile No.

Telephone No.

INTERNATIONAL SEARCH REPORT

International application No.

PCT/JP03/16982

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
-----------	--	-----------------------

A

JP 2002-358309 A (Hitachi Software Engineering Co., Ltd.),
13 December, 2002 (13.12.02),
& US 2002/184201 A1

1-4

A. 発明の属する分野の分類 (国際特許分類 (IPC))

Int. Cl⁷ G06F19/00

B. 調査を行った分野

調査を行った最小限資料 (国際特許分類 (IPC))

Int. Cl⁷ G06F19/00

最小限資料以外の資料で調査を行った分野に含まれるもの

日本国実用新案公報	1922-1996年
日本国公開実用新案公報	1971-2004年
日本国登録実用新案公報	1994-2004年
日本国実用新案登録公報	1996-2004年

国際調査で使用した電子データベース (データベースの名称、調査に使用した用語)

JICSTファイル (JOIS), WPI, INSPEC (DIALOG)

C. 関連すると認められる文献

引用文献の カテゴリー*	引用文献名 及び一部の箇所が関連するときは、その関連する箇所の表示	関連する 請求の範囲の番号
A	YONA, G. LEVITT, M. 'Within the Twilight Zone: A Sensitive Profile-Profile Comparison Tool Based Information Theory', Journal of Molecular Biology, 1 February 2002, Vol. 315, Issue 5, p. 1257-1275, especially p. 1259-1260 [on line] [retrieved on 29 January 2004], Retrieved from: <URL=http://www.cs.cornell.edu/golan/Papers/jmb02.pdf>	1-4

☒ C欄の続きにも文献が列挙されている。☐ パテントファミリーに関する別紙を参照。

* 引用文献のカテゴリー

「A」 特に関連のある文献ではなく、一般的技術水準を示すもの
「E」 国際出願日前の出願または特許であるが、国際出願日以後に公表されたもの
「L」 優先権主張に疑義を提起する文献又は他の文献の発行日若しくは他の特別な理由を確立するために引用する文献 (理由を付す)
「O」 口頭による開示、使用、展示等に言及する文献
「P」 国際出願日前で、かつ優先権の主張の基礎となる出願

の日の後に公表された文献

「T」 国際出願日又は優先日後に公表された文献であって出願と矛盾するものではなく、発明の原理又は理論の理解のために引用するもの

「X」 特に関連のある文献であって、当該文献のみで発明の新規性又は進歩性がないと考えられるもの

「Y」 特に関連のある文献であって、当該文献と他の1以上の文献との、当業者にとって自明である組合せによって進歩性がないと考えられるもの

「&」 同一パテントファミリー文献

国際調査を完了した日

30.01.2004

国際調査報告の発送日

10.2.2004

国際調査機関の名称及びあて先

日本国特許庁 (ISA/JP)

郵便番号100-8915

東京都千代田区霞が関三丁目4番3号

特許庁審査官 (権限のある職員)

高瀬 勤

5M

9069

電話番号 03-3581-1101 内線 3597

C (続き) . 関連すると認められる文献

引用文献の カテゴリー*	引用文献名 及び一部の箇所が関連するときは、その関連する箇所の表示	関連する 請求の範囲の番号
A	RYCHLEWSKI, L. et al. 'Comparison of sequence profiles. Strategies for structural predictions using sequence information', Protein Science, February 2000, Vol. 9. Issue 2, p. 232-241[on line][retrieved on 29 January 2004], retrieved from< http://www.proteinscience.org/cgi/reprint/9/2/232.pdf >	1 - 4
A	JP 2002-358309 A(日立ソフトウェアエンジニアリング)2002. 12. 13 &US 2002/184201 A1	1 - 4